

Monday Topic List Material

The enclosed material is intended to guide Monday interest group discussions. It has been organized thanks to Peter Wittenberg largely through his capacity as head of the Global Consortium.

The RDA Planning meeting organizing committee hopes this information and organization into a few breakouts will serve to put the right people in the right rooms so that discussion towards forming RDA working groups with clear value propositions can begin.

Topic List Monday

Overview

Data Foundation

- Foundation of Data Objects (2-Peter)
- Collection Properties (13-Reagan)
- Metadata Principles (3-Bill/Peter)
- Policy Rules (8-Reagan)
- Content Interoperability (14-Gerhard)

Registries

- PID Information Types (5-Daan)
- Data Type Registry (6-Larry)
- Center Registry (10-Stefan)
- Overcoming Data Friction (18-Ian)
- Data as a Service (7-Larry/Daan)
- Registry Interchange (19-Andrew)

Various

- Certification (4-Ingrid)
- Data-Publication Linking (15-Natalia)
- IPR Issues (16-Natalia)
- Marine Data Man (9-Helen)

Short 5 min presentations incl. goal – short discussion/recommendation

Topic Suggestions for Tuesday/Wednesday Overview

List of topics not covered on Monday

- Data Objects (1-Philippe)
- Urban Sciences Data (17-Charlie)

less clear scope

Charlie will arrive late at Monday

Short 5 min presentations incl. goal – short discussion/recommendation

Data Foundation Notes

- Foundation of Data Objects (2-Peter)
- Collection Properties (13-Reagan)
- Metadata Principles (3-Bill/Peter)
- Policy Rules (8-Reagan)
- Content Interoperability (14-Gerhard)

(2) Data Foundation

- Author: Stan Ahalt, Daan Broeder, Stefan Heinzl, Larry Lannom, Michael Lautenschlager, Reagan Moore, Arcot Rajasekar, Peter Wittenburg
- Problem:
 - crucial terms such as Data Object, Collections, PIDs, metadata, ext. and internal properties, policies and executable rules are not harmonized
 - basic IT principles (register syntax and semantics etc.) are not obvious
- Good Experience:
 - some projects have tried to properly define all terms and use IT principles
- Goal for WG:
 - harmonize concepts and terminology
 - create a generic cross-disciplinary reference model

(13) Science Collection properties

- Author: Arcot Rajasekar, Reagan Moore, Peter Wittenburg
- Problem:
 - very often lack contextual information for data collections to retest, reuse and repurpose data
 - there is a lack of agreement about essential properties which is often required for processing and assessing quality
 - lack agreements on essential domain-specific properties such as provenance, physical approximations and semantics
- Issue for WG:
 - agree on relevant properties and define and register them
 - define generic procedures to verify properties
 - define domain specific procedures to verify specific properties
 - define and create a starter kit

(3) Metadata and Data Architecture

- Author: Daan Broeder, Bill Michener, Jane Greenberg
- Problem:
 - metadata use is heterogeneous in different disciplines and projects
 - however, there are discipline independent aspects of MD that could be harmonized e.g.: *referencing relation with data and PIDs, DOs with multiple MD records, PID's tightly coupled MD, relation of MD with a DO's life cycle aspects*
- Good Experience:
 - Data community have modeling experience for self contained systems: repository systems, CMSs and also all-encompassing infrastructures like W3C.
- Issue for WG:
 - discuss metadata as part of the data architecture and identify common aspects.
 - inventory of the place of metadata in the most widely used data management systems.
 - make recommendations to harmonize these common aspects

(8) Policies

- Author: Ilkay Altintas, Irene Barg, Reagan Moore, et al.
- Problem:
 - the sheer amount of data will require automated operations on data objects/collections for various purposes (management, curation, etc)
 - assessment of quality can only be done if there are declarative statements that can be checked (thus formal policies turned into executable rules)
- Good Experience:
 - some repositories turn stepwise to using formalized policy rules using iRODS as execution engine
- Issue for WG:
 - define principles for policies, rules and execution frameworks
 - define areas to apply policies and sets of useful policies
 - define principles for quality assessments
 - define basic set of policies to be applied by trusted repositories

(14) Content Interoperability Approaches

- Author: Gerhard Budin, Frank van Harmelen, Andrew Maffei, Menzo Windhouwer, Peter Wittenburg
- Problem:
 - semantic bridges are hard and time intensive to create
 - there are different approaches (complex ontologies, taxonomies, concept registries, vocabularies, etc) that facilitate building pragmatic bridges
 - yet the knowledge about the various techniques is limited
 - there are many knowledge components around no one knows of
- Issue for WG:
 - a first BoF needs to understand the domain of semantic mapping and what has been done already, what the approaches are and which seem to have a potential for data practitioners
 - identify first simple steps to improve the situation and make this work more efficient
 - it is not meant to redo the W3C work with RDF, RDFS, OWL etc.

Registry Notes

- PID Information Types (5-Daan)
- Data Type Registry (6-Larry)
- Center Registry (10-Stefan)
- Overcoming Data Friction (18-Ian)
- Data as a Service (7-Larry/Daan)
- Registry Interchange (19-Andrew)

(5) Harmonization of PID Information Types

- Author: Daan Broeder, Michael Lautenschlager, Reinhard Budich, Ulrich Schwardmann, Larry Lannom, Maurice Bouwhuis, Pieter van Beek
- Problem:
 - the challenge of BIG data requires new strategies to automatically work with data objects – machines need to find useful information
 - important properties of data objects such as its checksum to allow verifying identity and integrity need to be associated with PIDs and found automatically independent of the PID system
 - yet there is no agreement on the information associated with PIDs
- Issue for WG:
 - we urgently need an agreed initial list of information to be associated with PIDs without overloading the resolution systems
 - we urgently need an open registry of such types with proper definitions including labels to be used
 - existing PID system APIs need to be adapted to support these types

(6) Type Registry

- Author: Larry Lannom, Daan Broeder
- Problem:
 - for many researchers it would be an enormous improvement and time saver if there would be a global registry of data types associated with an actionable reference that would allow to render the information
 - the simple MIME type system is not appropriate anymore since it does for example not support scientifically relevant complex objects
- Issue for WG:
 - compose a set of use cases for data type use and management
 - formulate a useful data model
 - design a functional specification for type registries
 - propose a federation strategy to allow multiple registries

(10) Data Center Registry

- Author: Stefan Heinzl, Johannes Reetz, Morris Riedel et al
- Author: Stefan Heinzl, Johannes Reetz, Morris Riedel et al
- Problem:
 - lack trusted global registries for various aspects as fixed points in a distributed environment with actors typically acting anonymously
 - lack a trusted global registry for repositories with machine (and human) readable information for automatic processing
- Experiences:
 - Grid community worked on it, there exist solutions (e.g. based on OGSA GLUE, DMFT CIM)
 - Community-specific solutions, e.g. IVOA registries, registry of registries
 - e.g. EUDAT has community and operational requirements on a site registry
- Issue for WG:
 - how to set up a world wide registry for data centers (administrative domains), e.g. operating repositories and other data services
 - which attributes do we need and how to define a flexible schema
 - how to register agreed attributes
 - work out a realization plan, render corresponding information models

(18) Overcoming Data Friction via Outsourcing and Process Automation

- Author: Ian Foster
- Problem:
 - in many small labs with long tail data efforts and costs for proper data management are immense
 - consequence is loss of data of many researchers
- Good Experience:
 - Flickr, Dropbox, etc. are good examples for reducing part of this effort
 - they offer data as a service
- Issue for WG:
 - how to design frameworks and services open to all researchers?
(upload, manage, store, preserve, curate, discover, share, annotate, etc.)
 - how to incentivize such services that enhance value of data?

(7) Data as a Service

- Author: Herman Stehouwer, Dieter van Uytvanck, Peter Wittenburg, Larry Lannom
- Problem:
 - most people (researchers, citizens) are unable to cope with the data itself, but want to have simple services on data
 - need a registry of services on data that can be used
- Good Experience:
 - Google and others demonstrate how popular such services on data are
- Issue for WG:
 - define metadata principles to describe services and to register them to make them discoverable
 - define a global registry system
 - define automatic profile matching to help users

(19) Registry Interchange

- Author: Andrew Treloar
- Problem:
 - discipline discovery services and their contents are not easily discoverable by people outside the communities they serve
- Good Experience:
 - <http://researchdata.ands.org.au/> is solving this problem for Australia, but the problem is international and the solution needs to be as well
- Issues for WG:
 - develop simple, extensible, and flexible models to enable a global network of data discovery
 - pilot some reference/ proof of concept registry interchanges amongst two or more of the working group participants
 - scope any lightweight international infrastructure needed to scale the model

Other Notes

- Certification (4-Ingrid)
- Data-Publication Linking (15-Natalia)
- IPR Issues (16-Natalia)
- Marine Data Management (9-Helen)

(4) Certification of Digital Repositories

- Author: Ingrid Dillo, Peter Doorn
- Problem:
 - confronted with a growing need to share data
 - access to data is widely anonymous by researchers who don't know each other
 - thus need measures to establish trust – one is to guarantee trusted repositories by assessing the quality of their processes
- Good Experience:
 - quality guidelines for trustworthy repositories available already (RAC/ISO, DSA, DIN, etc.)
- Goal for WG:
 - how can we implement and fund assessment practices worldwide?
 - how do we raise awareness that repositories need to be certified?

(15) Data-Publication Linking

- Natalia Manola, Najla Rettberg, Birgit Schmidt, Oya Yildirim Rieger, Marten Hoogerwerf, Jochen Schirrwagen, Norbert Losssau, Donatella Castelli, Alicia López Medina, Sarah Callaghan
- Problem
 - scientists have different notions of enhanced publications
 - lack of systematic and machine-readable information package for text-based publication and underlying research data
 - links between articles and data are already being made and the real challenge is to do it in a consistent (cross-discipline) and durable way
- Good Experience
 - prototype work by OpenAIRE with diverse data communities EBI (life sciences), DANS (social sciences), BADC (climate data)
- Issues for WG
 - types of data to be linked (e.g., visualizations, datasets, entries in large databases) and their role to the textual publication (e.g., supplementary files, research data)
 - types of relationships of linked data (*associations*),
 - scientific collections of data (*aggregations*) representation
 - versioning issues to capture changes made over time
 - responsibility: who defines, who maintains and who publishes the links
 - durability: stable, long-term network of related publications and datasets

(16) IPR Frameworks

- Andreas Wiebe, Paul F. Uhlig, Peter Murray-Rust, Enrique Alonso García, Natalia Manola, Alicia López Medina, Norbert Lossau
- Problem
 - research data are usually not covered by copyright
 - access to and distribution of information very often based on contractual arrangements
 - lack of shared intellectual property rights (IPR) framework
 - access, distribution, and improvement of data is influenced by cultural factors
 - cross-country, cross-infrastructure and cross-discipline IPR issues
 - different (language) IPR framework representations
- Good Experience
 - OpenAIRE task “Study on licensing of publications and research data”
 - COAR working group on OA repositories for publications
 - DCC (Data Curation Centre in UK) - how to guides, best practices
 - the Open Digital Rights Language - <http://www.w3.org/community/odrl/>
- Issues for WG
 - develop case based scenarios as a basis for a legal analysis
 - analyse different types of usage and enrichment
 - bring communities up to speed with existing representation languages for IPR frameworks

(9) Marine Data Management

- Author: Helen Graves
- Problem: Use of a variety of standards, formats, co-ordinate systems and best practice act as a barrier to the development of a common marine data management infrastructure on a global scale
- Good Experience:
Regional initiatives in Europe, the USA and Australia have implemented distributed marine data management infrastructures e.g. SeaDataNet, Geo-Seas, IOOS, the Australian Ocean Portal, the Rolling Deck to Repository (R2R) and the IODE Ocean Data Portal.

Goal for WG:

- Review existing standards, formats, vocabularies and best practices
- Develop a consensus on those that should be adopted as the 'community' standards'
- Document and disseminate these agreed 'standards'
- Promote the adoption of this common approach to marine data management

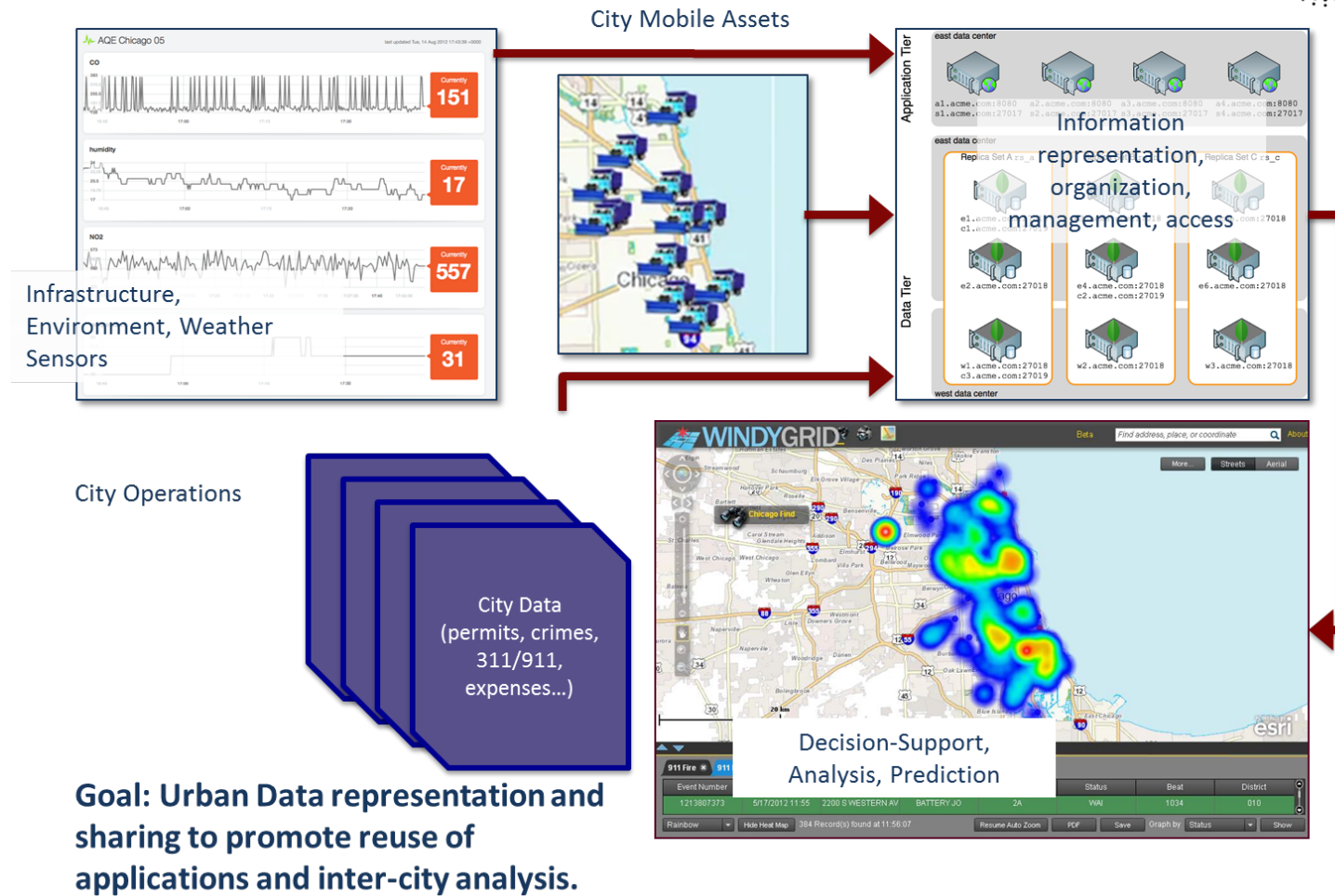
Notes about Remaining Topics

- Urban Sciences Data (17-Charlie)
- Data Objects (1 – Philippe Journeau)

(17) Urban Sciences Data

- Author: Charlie Catlett

Data-Driven Urban Design and Optimization



(1) Data Objects

- Author: Philippe Journeau
- Problem: Meaning, Nature of Data Objects and versus Objects have to be enlightened for at least the following strategic issues:
 - data management, storage, prioritization, preservation
 - expectation that data will be linkable rather than the contrary (is the related curve going to converge or diverge)
 - classical linguistic and moreover knowledge relationship between (formal) data and (read) world and more widely meaning and objects, become industrial and macroeconomic issue for the next types of web(s)
- Good Experience:
- Goal for WG: Produce a landscape of Nature of Data versus other Objects

Notes Projects/Institutions

- not real topics -

- HEP Data Man (11-Jamie)
- Life and Med Science DM (12-Andrew)

(11) DM in High Energy Physics

- Author: Jamie Shiers

Data at 100-1000PB scale; globally distributed; super-linear growth

Driver is production service to LHC experiments;

Tier model (0/1/2); impressive service level across O(100) sites;

Single source for “real” data; many for simulated;

Moving from static to dynamic data placement;

Effective use of multiple 10Gbit links [100Gbit inter-T0];

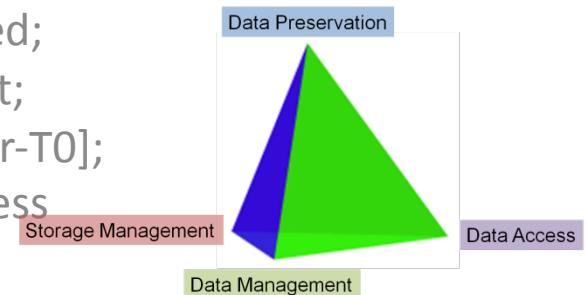
Federated data stores across WAN; remote access

Existing partnerships with other disciplines

Support today LS, ES, A&A projects – more in future

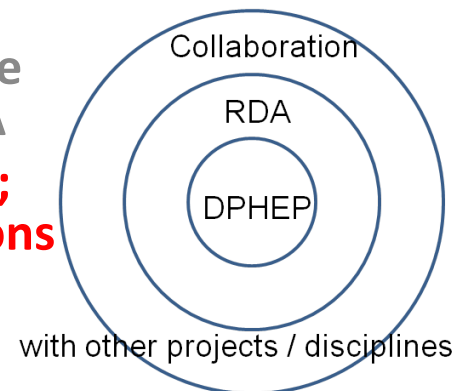
Key activities: Long-term Data Preservation (DPHEP)

E.g. preserving ability to re-analyse LHC data in ILC era (~2030)



Bottom line: facing in production today many of the challenges we believe are being addressed by RDA

Our goal: share experience; collaborate effectively; move towards more “standard” / sustainable solutions



(12) DM in Life and Medical Sciences

- Author: Andrew Lyall
- Proposal for a Topic to be considered at the Research Data Alliance Global Data Meeting to be held in Washington on 1 October 2012
- The Human Genome Project was the archetype for large multinational big-science projects in biology
- It has created new ways of doing biology and medical research where data generation and analysis are the main tasks
- Hospitals and other health-research institutes will be generating and using huge amounts of data (cf. ELSI)
- The same applies to research in agriculture, biotechnology etc...
- Acquiring, storing, archiving and moving these data around are now significant challenges
- Harmonisation of these activities will be important in order to use these data to tackle the grand challenges: healthcare for aging populations, food security, environmental protection, etc
- 18 Month deliverable – a catalogue of all large generators and users of “omics” data, to include details of standards, ontologies, data analysis pipelines and data management systems